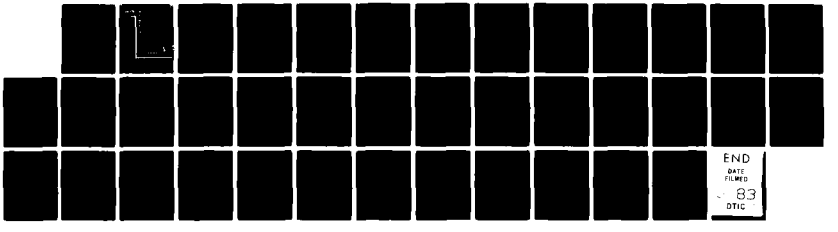
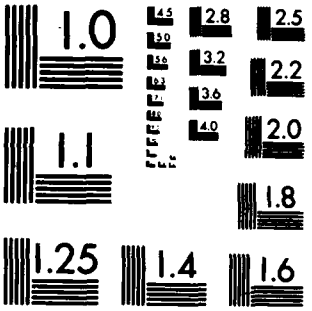
AD-A123 987 REGRX: A COMPUTERIZED STEPWISE REGRESSION ALGORITHM WITH RESIDUAL ANALYSIS(U) AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX W G ALBERT ET AL. FEB 83 AFHRL-TP-82-40 1/1

UNCLASSIFIED F/G 9/2 NL



END
DATE FILMED
83
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFHRL-TP-82-40

(12)

AIR FORCE HUMAN RESOURCES LABORATORY

AD A123987

REGRX: A COMPUTERIZED STEPWISE REGRESSION ALGORITHM WITH RESIDUAL ANALYSIS

By

Walter G. Albert
Janos B. Koplyay
Larry K. Whitehead

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

February 1983

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Manpower and Personnel Division

J. P. AMOR, Lt Colonel, USAF
Chief, Manpower and Personnel Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFHRL-TP-82-40	2. GOVT ACCESSION NO. AD A123987
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)* REGRX: A COMPUTERIZED STEPWISE REGRESSION ALGORITHM WITH RESIDUAL ANALYSIS	5. TYPE OF REPORT & PERIOD COVERED Final
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Walter G. Albert Janos B. Koplyay Larry K. Whitehead	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Manpower and Personnel Division Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62703F 77192201
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235	12. REPORT DATE February 1983
	13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS *(of this report)* Unclassified
	15.a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of this abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

algorithm
mathematical methodology
regression
residual analysis
statistical methodology
stepwise regression

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

REGRX, a stepwise regression algorithm with residual analysis, functions as a subroutine in the computer-based TRICOR utility correlation and regression system. The comprehensive statistical/mathematical methodology comprising REGRX has served as a powerful tool in the development of Air Force prediction systems.

This paper contains the technical details that are necessary for the user to take complete advantage of the analytical capabilities of REGRX. This information includes an in-depth discussion of the REGRX mathematical/statistical methodology with associated computational formulas and printed output samples. Therefore, this document aids researchers in executing and interpreting the results of the REGRX subroutine.

DD Form 1473 EDITION OF 1 NOV 65 IS OBSOLETE
1 Jan 73

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Date Entered)*

AFHRL Technical Paper 82-40 February 1983

REGRX: A COMPUTERIZED STEPWISE REGRESSION ALGORITHM WITH RESIDUAL ANALYSIS

By

Walter G. Albert
Janos B. Koplyay
Larry K. Whitehead

Reviewed by

C. Deene Gott
Chief, Experimental Design and Analysis Function
Manpower and Personnel Division

Submitted for publication by

Janos B. Koplyay
Chief, Manpower and Force Management Systems Branch
Manpower and Personnel Division

This publication is primarily a working paper.
It is published solely to document work performed.

PREFACE

The research was completed under Project 7719, Force Acquisition and Distribution System, Task 771922, Development of Analytic Methodology for Air Force Personnel Research. Dr. Janos B. Koplyay, Mr. Larry K. Whitehead, and Mr. William S. Mathon were primarily responsible for the development of the REGRX methodology and its implementation on the AFHRL UNIVAC 1108 computer system. Mr. Stephen D. Poole is due special acknowledgment for incorporating the REGRX methodology into the TRICOR utility correlation and regression system.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

REGRX: A COMPUTERIZED STEPWISE REGRESSION ALGORITHM WITH RESIDUAL ANALYSIS

I. INTRODUCTION

Regression analysis programs are commonly used to develop prediction systems for Air Force personnel research, e.g., a system for accurately predicting future job performance of enlisted individuals based on information obtained during their Air Force careers. This information might include such predictor variables as aptitude or ability test scores, biographical data, and physical attributes. Requirements for the development of such prediction systems within the technical programs of the Air Force Human Resources Laboratory (AFHRL) are numerous.

Prior to implementation of the UNIVAC 1108 computer system at AFHRL, the majority of regression analyses were accomplished by the REGRED single correction iterative algorithm (Ward, Hall, & Buchhorn, 1967). This algorithm had two major disadvantages: (a) it might not converge if two or more variables were highly intercorrelated, and (b) since it did not identify redundant variables, an incorrect number of degrees of freedom could be used in calculating the F-ratio discussed in Section IV. The convergence problem was eliminated by a modified interative algorithm called REGREF (Ward et al., 1967), which corrected on three weights per iteration simultaneously; however, the REGREF algorithm still failed to identify redundant variables.

During the conversion from the previous computer system to the UNIVAC 1108 computer system, a computerized regression algorithm, REGRX, specifically tailored to the requirements of analyses performed by laboratory task scientists was developed to exploit the capabilities of the UNIVAC 1108. REGRX was implemented to improve the laboratory's problem-solving capabilities by allowing for

identification of redundant predictor variables, an exact solution at each step of the algorithm, extensive residual analysis, forcing certain predictor variables into the final equation, and direct generation of transformed predictor variables.

Shortly after the REGRX program had been implemented and thoroughly tested on the UNIVAC 1108, the algorithm was incorporated as a subroutine, REGRX, into the TRICOR utility correlation and regression software package, which immediately resulted in improved analytical capabilities and product quality. Since that time, the REGRX subroutine system has undergone several modifications, and has now been implemented on the UNIVAC 1100/81.

The purpose of this paper is to acquaint the potential user with the current capabilities of REGRX. Technical details are discussed to enable the user to take complete advantage of the analytical capabilities of REGRX. This information includes a brief introduction to the stepwise regression technique, an in-depth discussion of the REGRX algorithm, a comprehensive listing of the computational formulas and definitions of the resulting statistics, and a description of the algorithm's residual analysis facilities. Specific details for running the TRICOR software package on the UNIVAC 1100/81 are available at AFHRL in an automated users manual titled TRICOR: Utility Correlation and Regression System.

II. STEPWISE REGRESSION AND MODEL BUILDING

The REGRX regression procedure is a stepwise augmentation and elimination algorithm. The stepwise technique (Dixon, 1968; Draper & Smith, 1966; Efroymson, 1960; Pope & Webster, 1972; Goldberger, 1961; Goldberger & Jochems, 1961) is used primarily as a research tool to aid in the screening and selection of variables in the development of a mathematical model of a statistical relationship

between a response and a set of independent variables. It is usually desirable that a model of the response-independent variable relationship contain as few independent variables as possible; therefore, for those cases in which a large number of variables are identified as having some influence on the response, it is necessary that some form of variable selection be performed.

The stepwise algorithm is a systematic process for adding variables to or deleting variables from a given initial linear model. First, the response variable is regressed on the set of independent variables comprising the initial model. At each subsequent step, a new regression equation is derived from the equation at the previous step either by deleting a variable for which the partial F-statistic testing for a zero coefficient falls below a preassigned value or by adding a variable for which the partial F exceeds a preassigned value. At some point, this process of adding and deleting variables is interrupted, and the variables in the final regression equation are taken as the components of a new model. Dixon (1968) provides a more complete description of how the stepwise procedure may be incorporated into a model building program. The REGRX stepwise algorithm is discussed in more detail in the following paragraphs. In addition, Appendix A provides a general summary of the correlation approach to regression, and Appendix B provides the computational details of the REGRX algorithm.

III. DESCRIPTION OF REGRX ALGORITHM

At each step of the algorithm, the independent variables are divided into two sets, L and E. L is the set of variables in the regression equation for the current step. Set E contains all independent variables that are not contained in L. Thus, when a variable is added to L, it is simultaneously deleted from E and vice versa.

Initially, set L does not contain any variables and set E may contain a set of "forced" variables that have been designated by the task scientist to appear in all calculated regression equations. If no variables are designated as "forced," the first variable to be added to L is the independent variable most highly correlated with the dependent variable. If set E contains "forced" variables, the first variable to be added to L is the "forced" variable most highly correlated with the dependent variable; the other "forced" variables are considered for addition to L by the stepwise procedure before the remaining E variables. The set of "forced" variables that are added to L are denoted by F. A given variable designated as "forced" is not allowed to be an element of L if the squared multiple correlation coefficient for the regression of the given variable on the set L is greater than or equal to 1.0 - TOL where the value of TOL can range from 10^{-1} to 10^{-8}. The user should specify a set of independent variables as "forced" if the regression problem requires that these predictors be present in the final regression equation.

At each subsequent step, the stepwise procedure regresses the dependent variable and each variable in E on the variables in L, and one of the following outcomes occurs:

1. A variable in L - F (the set of variables remaining in L after the variables in F have been removed from consideration) is deleted from L if the partial F-statistic testing for a zero coefficient is less than a preassigned value and if no other variable in L - F has a smaller partial F.

2. A variable in E is added to L if (a) no variable in L - F satisfies the removal criterion; (b) the squared multiple correlation coefficient for the regression of the added variable on the set L is less than 1.0 - TOL; (c) after adding the variable to L, the partial F-statistic testing for a zero coefficient exceeds a

preassigned value; and (d) no other variable in E satisfies (b) and has a larger partial F.

3. If neither of the preceding outcomes occurs or if the number of steps exceeds a preassigned number, then the stepwise procedure terminates.

IV. REGRESSION OUTPUT WITH COMPUTATIONAL FORMULAS AND COMMENTS

The following information describes all of the printed output that can be generated by the REGRX subroutine system except for the residual plots described in Section V. Subsections A to F are output produced for each step printed. Subsections G to K are optional output. Items appearing in upper-case letters are presented exactly as they appear in the printed output. Details on printing options available are stated in an automated users manual titled TRICOR: Utility Correlation and Regression System.

A. MULTIPLE RSQ: Coefficient of simple determination between the predicted scores and the observed values for the dependent variable. If a is the regression constant and b_j is the estimated regression coefficient for the j^{th} variable, then the k^{th} predicted score is

$$\hat{y}_k = a + \sum_{j=1}^{p} b_j x_{jk}$$

where x_{jk} is the k^{th} observation of variable j and p is the number of predictors in the prediction equation.

B. ANALYSIS OF VARIANCE

1. Regression Degrees of Freedom: $DF_{Reg} = p_i$ = number of predictors in step i

2. Regression Sum of Squares: $SS_{Reg} = \sum_{k=1}^{n} (\hat{y}_{ik} - m_{\hat{y}i})^2$

where n = number of observations

$\hat{y}_{ik}$ = predicted score of k^{th} observation in step i

$m_{\hat{y}i}$ = mean predicted score in step i

3. Regression Mean Square: $MS_{Reg} = SS_{Reg}/DF_{Reg}$

4. Residual Degrees of Freedom: $DF_{Res} = n - p_i - 1$

5. Residual Sum of Squares: $SS_{Res} = \sum_{k=1}^{n} r^2_{ik}$

where $r_{ik} = y_{ik} - \hat{y}_{ik}$ = residual of k^{th} observation in step i

6. Residual Mean Square: $MS_{Res} = SS_{Res}/DF_{Res}$

7. F-Ratio = (Regression Mean Square)/(Residual Mean Square)

C. STD ERR EST = $\sqrt{MS_{Res}}$

D. REG CONST: Estimate of the mean response when all of the predictors have a value of zero.

E. VAR: Variable (ID and name) to enter the prediction system in step i

F. Prediction System Table

1. VARIABLE: Variable ID and name for each predictor in the system in step i

2. REGRESSION WEIGHT (b_{ij}): Estimates of the regression parameters β_{ij} for variable j in step i which indicate the change in the mean response associated with a unit change in the corresponding predictor variable when all other predictor variables are held constant

3. STANDARD WEIGHT (B_{ij}): If SD_y is the standard deviation of the dependent variable and SD_j is the standard deviation of variable j, then

$$B_{ij} = b_{ij} \frac{SD_j}{SD_y}$$

4. SQ CORRELATION VARIABLE VS REST (R^2_{ij}): The squared multiple correlation coefficient for the regression of variable j on all of the other predictors in the prediction system in step i

5. STANDARD DEVIATION OF REGRESSION WEIGHTS (SD_{wj}):

$$SD_{wj} = \sqrt{\frac{MS_{Res}}{(SD_j)^2 \ (n-1) \ (1-R^2_{ij})}}$$

If variable j is uncorrelated with the other predictors, then

$$SD_{wj} = \sqrt{\frac{MS_{Res}}{(SD_j)^2 \ (n-1)}}$$

Note that this is the smallest value that SD_{wj} can assume.

6. INDEPENDENT CONTRIBUTION (ΔR^2_{ij}): Amount by which R^2_i would decrease if variable j were removed from the prediction system

$$\Delta R^2_{ij} = \frac{1}{(n-1)} \frac{MS_{Res} \ b^2_{ij}}{(SD_y)^2 \ (SD_{wj})^2} = \frac{1}{(n-1)} \frac{MS_{Res}}{(SD_y)^2} F_{ij}$$

where F_{ij} is the partial F-statistic for testing the null hypothesis that the j^{th} partial regression coefficient at step i equals zero

7. SQUARED PARTIAL CORRELATION $\left(r^2_{yj \ . \text{ all other predictors}}\right)$: The marginal contribution of predictor j in the proportionate reduction in the variance of the dependent variable when all of the other predictors have already been included in the prediction system

$$r^2_{yj \ . \text{ all other predictors}} = \frac{\Delta R^2_{ij}}{1 - (R^2_i - \Delta R^2_{ij})}$$

where R^2_i is the squared multiple correlation coefficient for the regression of the response variable on all of the predictors for step i

G. REGRESSION SUMMARY TABLE

For each step, the summary table gives the step number, the ID of the variable entered or removed, the coefficient of multiple correlation and coefficient of multiple determination for the regression of the dependent variable on all of the predictors in the prediction system, the change in the coefficient of multiple determination from the previous step, the residual mean square, the square root of the residual mean square, the F-ratio, the partial F value, and the number of predictors in the prediction system.

The formula used for computation of the F-ratio for step i is

$$F_i = \left(\frac{n-p_i-1}{p_i}\right)\left(\frac{R_i^2}{1-R_i^2}\right)$$

The partial F-ratio is directly related to the independent contribution for the variable entered at each step. An alternate computational formula for this statistic is

$$F_{ij} = \frac{(R_i^2 - R_{i-1}^2)/df_1}{(1-R_{max}^2)/df_2}$$

where R_i^2 = coefficient of multiple determination at step i

R_{i-1}^2 = coefficient of multiple determination at step i-1

R_{max}^2 = the larger of R_i^2 and R_{i-1}^2

df_1 = difference in the numbers of predictors in the prediction systems corresponding to steps i and i-1; consequently, df_1 will always equal 1

df_2 = n-p-1, with p being the number of predictors in the prediction system for the step corresponding to the larger of R_i^2 and R_{i-1}^2

This statistic is identical to the traditionally used comparison of "full model" versus "restricted model" in the iterative REGREF regression algorithm (Bottenberg & Ward, 1963).

H. LINEAR DEPENDENCIES

If the correlation matrix is not of full rank, i.e., some of the predictor variables are redundant, then the least squares normal equations will not have a unique solution. Least squares parameter estimates can still be obtained (Rao & Mitra, 1971; Searle, 1971); however, there will be infinitely many estimates, all equally good. An alternative is to identify redundancies and assign zero weights to the redundant variables, thereby eliminating them from the prediction system.

At each step, the REGRX algorithm computes a regression for each candidate entry variable on all of the variables in the prediction system. If the coefficient of multiple determination for any of these regressions is greater than 1 - TOL, where the value of TOL is specified by the user and ranges from 10^{-1} to 10^{-8}, the variable is considered redundant and will not be allowed to enter the prediction system. When a variable is identified as redundant in this way and at least one of the standardized partial regression coefficients is greater than or equal to 10^{-5}, the ID for the redundant variable is printed with the corresponding regression coefficients. The "intercept" printed is the regression constant for the prediction equation. A variable is also considered redundant if its entry into the prediction system would cause a linear dependency among those variables in the augmented system.

The value of TOL should be selected with care. Choice of an ideal value for TOL depends greatly on the data set to be analyzed. There may be a tendency to choose small values for TOL to allow as many variables as possible to enter the prediction

system. However, extensive testing has shown that known linear dependencies may not be identified by REGRX if TOL is set very low. Moreover, when the value of TOL is set at a small value, the following three undesirable situations are more likely to occur: (a) severe computational accuracy problems, (b) large standard errors for the regression coefficients, and (c) results being adversely affected by slight data recording errors. On the other hand, choosing a value for TOL that is too large may exclude predictor variables that are functions of other predictor variables even when those variables would contribute greatly to predictive efficiency. Unless specified otherwise, the value of TOL is set to 10^{-3}, which has been shown to be suitable for most applications.

I. RANGE TABLE

The Range Table gives the means, standard deviations, and maximum and minimum values for the observed values for the dependent variable, predicted scores, and residuals. The table also gives the residual variance and the coefficients of correlation and determination between the observed values for the dependent variable and the predicted scores, and between the residuals and the predicted scores.

The maximum and minimum range values for the observed values for the dependent variable and the predicted scores should be comparable. Vast differences between maximums or minimums, as compared to the residual standard deviation, may indicate either an error in the data or the inability to predict extreme values of the criterion, suggesting additional terms need to be included in the model.

J. TABLE OF LARGEST RESIDUALS

The table of largest residuals gives the case identification number, predicted score for the dependent variable, and residual and predictor values (includes variable name and identification

number for all variables in the prediction system) for the cases associated with the X largest residuals. X is the lesser of the following two quantities: 10, or the number of cases divided by 20. The detection of outliers or data errors is facilitated by printing extreme residual values. The range values for the residuals should be within plus or minus three standard deviations of the residual mean. This is in accord with the fact that in a normal population virtually all points lie within plus or minus three standard deviations of the mean.

K. TABLE OF RESIDUALS

This table is printed upon request by the user and lists for each observation, the case identification number, predicted score for the dependent variable, and residual value.

V. RESIDUAL PLOTS

Descriptions of the REGRX residual analysis facilities are presented below. To complement these descriptions, the reader is referred to Draper and Smith (1966) where an excellent discussion of the analysis of residuals is presented.

Plot of Residuals vs Predicted Scores

A plot of residuals versus predicted scores for a typical REGRX problem is shown in Figure 1. The residual axis appears vertically on the page. Two scales are given: (a) the standardized residual, and (b) the residual itself. The two rows immediately above and below the plot represent the predicted score axis. The first and last entries on this axis are the smallest and largest predicted scores. If r_j and s_r are the j^{th} residual and the residual standard deviation, respectively, then the standardized residual is r_j/s_r. When the residuals follow a normal distribution with variance s_r^2, the standardized residuals follow a normal distribution with variance unity. Thus, approximately 95% of the residuals would be expected to fall between -2 and +2 on this axis.

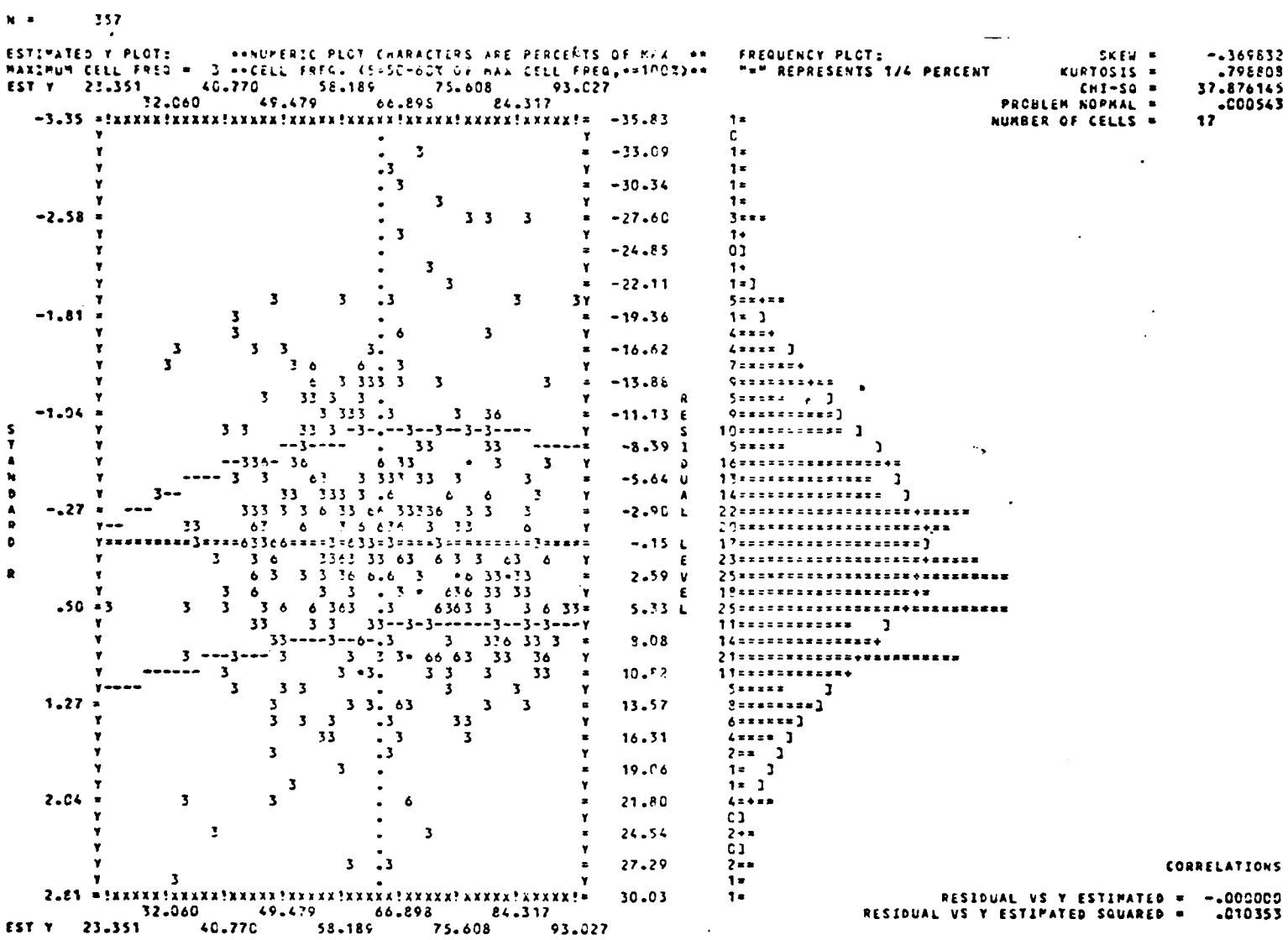


Figure 1. Plots of residuals versus predicted scores and residual frequency

A cell is defined as the intersection of a row and a column of the graph. Each cell may have several points plotted within it. The maximum cell frequency, i.e., the largest number of points plotted within a cell of the graph, is printed at the top of the plot. In Figure 1, MAXIMUM CELL FREQ = 3. The number of points plotted within a cell is indicated by a numeric character or asterisk. The numeric character L indicates that between 10L and 10(L+1) percent of the maximum cell frequency points were plotted in the respective cell. An asterisk indicates that the cell contains exactly MAXIMUM CELL FREQ points. For example, a numeric character 4 with a MAXIMUM CELL FREQ=20 would indicate that between 40 and 50 percent of 20 points, i.e., between 8 and 10 points, are plotted in that particular cell. Truncation is assumed. If MAXIMUM CELL FREQ = 5, then between 40 and 50 percent of 5 points, i.e., 2 points, are plotted in a cell where L=4. Similarly, a numeric character of 2 would indicate that exactly one point is plotted in the cell. The printing of cell frequency count indicators takes precedence over all other printing requirements.

A row of equal signs and a column of periods identify the residual mean and the predicted score mean, respectively. The REGRX algorithm performs a quadratic regression of the residuals on the predicted scores, i.e., the independent variable appears in the first and second degree in the model. The dashes in Figure 1 depict quadratic regressions of the residuals on the predicted scores for the set of points located above the initial quadratic regression and for the set of points located below the initial quadratic regression.

This plot is useful in detecting heteroscedasticity (unequal variances for error terms) and model inadequacies. If neither of these abnormalities is present, the plotted points should appear as a random scatter of points about a line parallel to the predicted score axis and intercepting the residual axis at zero. However, a systematic pattern of points such as a wedge (heteroscedasticity) or curvilinear (model inadequacy) shape signals the need for corrective

action. If heteroscedasticity is present, the analyst should consider the use of weighted least squares or various transformations of the dependent variable such as $\sqrt{Y}$, 1/Y, and log Y. Similarly, methods of dealing with model inadequacy are transforming the dependent variable or including additional terms in the model such as square or cross-product terms.

This plot is also useful in detecting outliers (points that are more than three standard deviations from the residual mean). Since the least squares fit is "pulled" disproportionately toward these observations, outliers should be carefully examined to determine if they convey important information about the analysis or if they resulted from a procedural error such as a miscalculation, inaccurate recording or equipment malfunction. In general, an outlier should not be eliminated from the analysis unless the task scientist can identify an error source causing the extreme value.

The plotted points in Figure 1 exhibit no severe abnormalities such as those exhibited in Draper and Smith (1966); however, the dashes do show a slight curvilinear trend in the data.

Residual Frequency Plot

This plot (shown in Figure 1) is printed on the same page as the previously discussed plot. The residual axis appears vertically on the page. The number of points falling within each interval of the residual axis is printed vertically in the left margin. If N is the total number of residuals, i.e., the total number of cases, and if L is the frequency count for an interval on the residual axis, then 100L/N percent of the residuals fall in this particular interval. Each equal sign represents one-fourth percent of the total number of points; therefore, 400L/N equal signs would be printed on the line corresponding to that interval.

A normal frequency curve is superimposed on the frequency plot as a series of plus signs and "]" symbols, the plus sign representing an

overlap between an equal sign and a "]" symbol. The purpose of this superimposed curve is to provide a visual standard against which the observed frequency curve can be compared. The chi-square value printed at the top of the plot provides a quantitative test of the hypothesis that the residuals are normally distributed. NUMBER OF CELLS is the number of intervals (ranges from 5 to 49 depending on sample size) used for this test. PROBLEM NORMAL is the probability of exceeding the chi-square value when the residuals are normally distributed. Measures of the asymmetry (skewness) and flatness (kurtosis) are printed above the chi-square value.

The correlation coefficients between the residuals and the predicted scores and between the residuals and the squared predicted scores are printed below the plot. The first of these correlation coefficients is expected to be zero and the second is an indicator of the quadratic tendency between the residuals and the predicted scores.

Cumulative Frequency Plot

The residual axis appears vertically on the page in Figure 2 in the same manner as for the plot of residuals versus predicted scores.

The horizontal axis at the top of the graph is the cumulative frequency axis with the cumulative frequencies given as fractions of the total sample size. Thus, if N=357 is the total number of cases, a fraction of X=.510 would indicate a cumulative frequency of (X)(N)=(.510)(357)=182. The frequency curve is plotted using the symbol "F" or an asterisk. A curve drawn through the Fs and asterisks should resemble a normal cumulative frequency curve. For visual comparison, a normal cumulative frequency curve could have been superimposed on the graph. However, it is easier to observe a deviation from a straight line than it is from the normal curve. Therefore, as an alternative, the observed frequency plot was transformed in such a manner that it gives a straight line if the original plot was in fact a normal cumulative frequency curve but

N = 357

CUMULATIVE FREQUENCY PLOT:

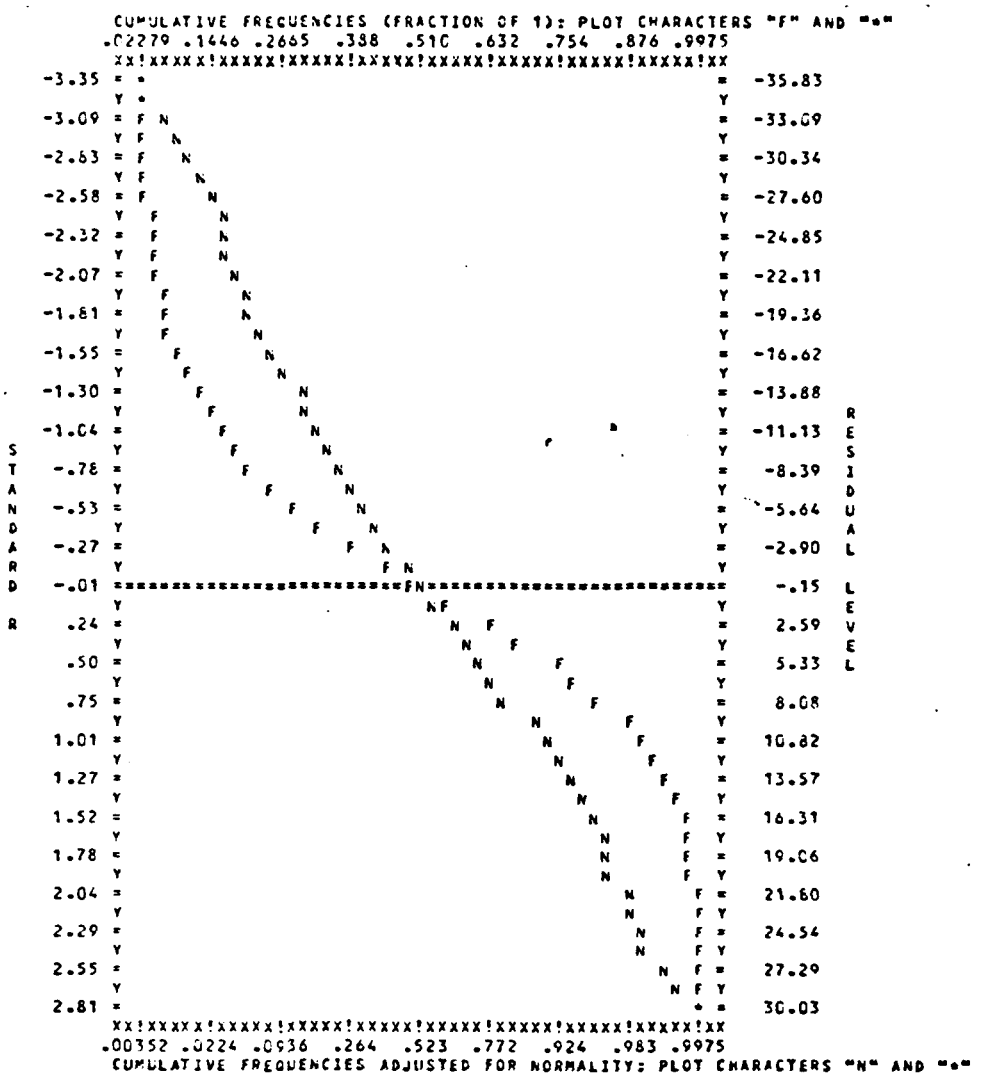


Figure 2. Cumulative frequency plot

would deviate from a straight line if the original differed from a normal curve.

The horizontal axis at the bottom of the graph is the cumulative frequency axis for the transformed curve. The transformed cumulative frequency curve is superimposed on the cumulative frequency plot using the symbol "N" or an asterisk. The asterisk represents a point at which the original curve and its transform intersect. The horizontal row of equal signs corresponds to the residual mean.

Plot of Residuals versus Predictor Values

A plot of the residuals versus a predictor variable is shown in Figure 3. The values on the horizontal axis are the values of the predictor. A horizontal row of equal signs and a vertical row of periods identify the residual mean and predictor variable mean, respectively. The correlation coefficients between the residuals and the predictor variable scores and between the residuals and the squared predictor variable scores are printed below the plot. The residual axis appears vertically on the page in the same manner as for the plot of residuals versus predicted scores. In addition, the maximum cell frequency value, numeric characters, and asterisk and dash symbols are printed as before.

This plot is useful in detecting heteroscedasticity and model inadequacies. As before, the absence of abnormalities is indicated by a random scatter of points about a line parallel to the predictor score axis intercepting the residual axis at zero. A wedge-shaped point scatter is indicative of heteroscedasticity. Possible corrective actions that should be investigated by the task scientist include the use of weighted least squares and various transformations of the dependent variable such as Y/X_j or $Y\sqrt{X_j}$. A curvilinear trend in the point scatter is indicative of an inadequate model. Possible corrective actions for this abnormality include the use of various transformations on the dependent variable and adding terms to the model such as square or interaction terms. Outliers are also easily

identified in these plots. The plotted points in Figure 3 exhibit no severe abnormalities.

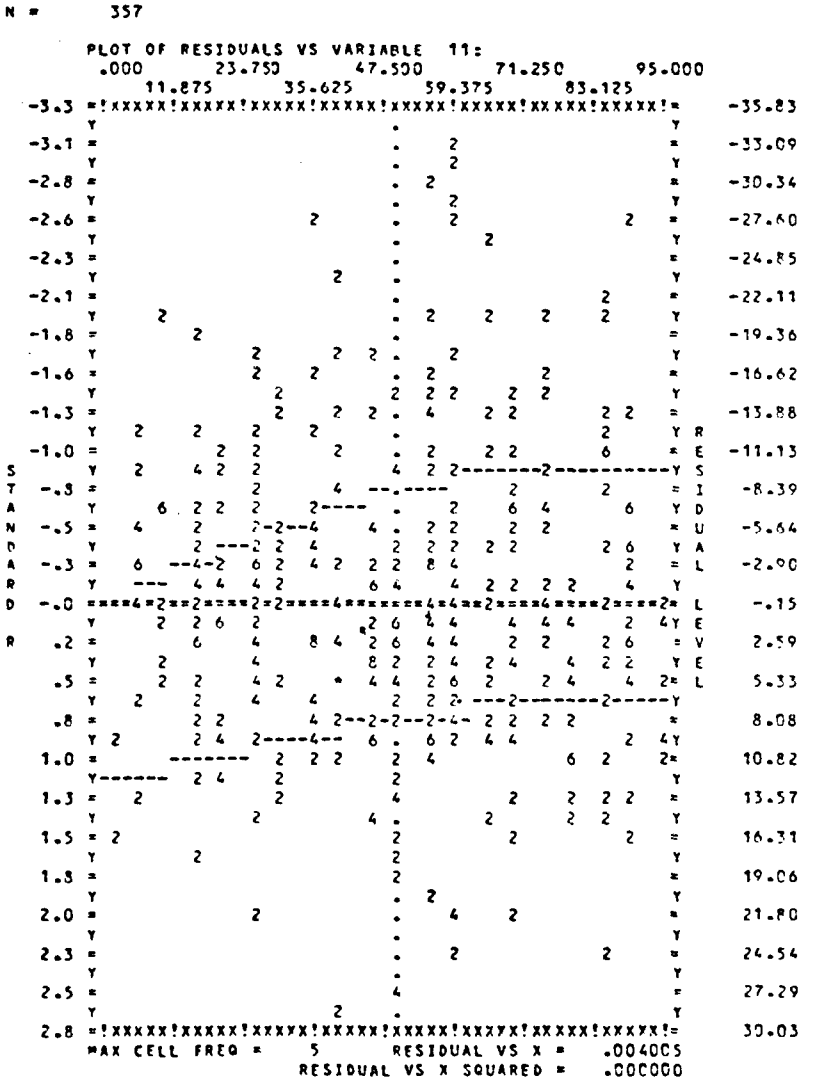


Figure 3. Plot of residuals versus predictor values

REFERENCES

Bottenberg, R. A., & Ward, J. H., Jr. Applied multiple linear regression. PRL-TDR-63-6, AD-413 128. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, March 1963.

Dixon, W. J. BMD: Biomedical computer programs. Berkeley, CA: University of California Press, 1968.

Draper, N. R., & Smith, H. Applied regression analysis. New York: Wiley, 1966.

Efroymson, M. A. Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), Mathematical methods for digital computers. New York: Wiley, 1960, 191-203.

Goldberger, A. S. Stepwise least squares: residual analysis and specification error. Journal of the American Statistical Association, 1961, 56, 998-1000.

Goldberger, A.S., & Jochems, D. B. Note on stepwise least squares. Journal of the American Statistical Association, 1961, 56, 105-110.

Pope, P. T., & Webster, J. T. The use of an F-statistic in stepwise regression procedures. Technometrics, 1972, 14, 327-340.

Rao, C. R., & Mitra, S. K. Generalized inverse of matrices and its applications. New York: Wiley, 1971.

Searle, S. R. Linear models. New York: Wiley, 1971.

Ward, J. H., Jr., Hall, K., & Buchhorn, J. PERSUB reference manual. PRL-TR-67-3 (II), AD-660 579. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, August 1967.

REFERENCE NOTE

Specific details for running the TRICOR software package on the AFHRL UNIVAC 1108 are available in an unpublished automated users manual at AFHRL titled TRICOR: Utility Correlation and Regression System.

APPENDIX A: CORRELATION APPROACH TO REGRESSION[1]

Regression methodology is concerned with the problem of estimating the parameters $\theta_1, \theta_2, \ldots, \theta_p$, and α in the linear model $\mathbf{Y} = \alpha\mathbf{1}+\mathbf{x}_1\theta_1+\ldots+\mathbf{x}_p\theta_p+\mathbf{E} = \alpha\mathbf{1}+\mathbf{X}\boldsymbol{\theta}+\mathbf{E}$. For this problem, the following assumptions are commonly made: $\mathbf{X}$ is a matrix of known form and the error component, $\mathbf{E}$, is assumed to be distributed with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}$. According to the Gauss-Markov Theorem, the minimum variance unbiased linear estimators $\hat{\alpha}, \hat{\theta}_1,\ldots,\hat{\theta}_p$ for the parameters $\alpha, \theta_1, \ldots, \theta_p$ are obtained by the method of least squares. This method leads to a system of linear equations, called the normal equations, which are solved for $\hat{\alpha}$ and $\hat{\boldsymbol{\theta}}$.

$$\mathbf{1}^T\mathbf{Y} = \mathbf{1}^T\mathbf{1}\hat{\alpha} + \mathbf{1}^T\mathbf{X}\hat{\boldsymbol{\theta}}$$
$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{1}\hat{\alpha} + \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\theta}}$$

The superscript T denotes that the columns of $\mathbf{X}^T$ are the rows of $\mathbf{X}$ and the rows of $\mathbf{X}^T$ are the columns of $\mathbf{X}$. To decrease the effects of rounding error in the computation of the solution of the normal equations, the observations X_{ij} and Y_j are first centered and then rescaled to standardized form z_{ij} and y_j, where

$$z_{ij} = (X_{ij} - \bar{X}_i)/s_i, \qquad y_j = (Y_j - \bar{Y})/s_y$$

X_{ij} = jth observation of variable i

$\bar{X}_i$ = sample mean for variable i

s_i = sample standard deviation for variable i

---

[1]Matrices, vectors, and scalars will be denoted by uppercase boldface letters, lowercase boldface letters, and upper or lowercase regular typeface, respectively. Numerically subscripted scalars identify elements of matrices (row identification, column identification) or vectors (row identification) and numerically subscripted matrices identify partitioned elements of matrices.

The normal equations can be rewritten in terms of the standardized variables as

$$\mathbf{g} = \mathbf{R}_{11}\mathbf{b}$$

where $\mathbf{g} = \frac{1}{n}\mathbf{Z}^T\mathbf{y}$

$\mathbf{Z} = (z_{ij})$

$\mathbf{R}_{11} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$

$\mathbf{b} = \frac{1}{s_y}\mathbf{S}\hat{\boldsymbol{\theta}}$

$\mathbf{S} = \text{diag}\ (s_i)$

$\mathbf{S} = \text{diag}\ (s_i)$ means that $\mathbf{S}$ is a diagonal matrix with the ith diagonal entry equal to s_i. The estimates $\hat{\boldsymbol{\theta}}$ are calculated by solving the system $\mathbf{R}_{11}\mathbf{b} = \mathbf{g}$ for $\mathbf{b}$ and then computing $\hat{\theta}_i = \frac{s_y}{s_i} b_i$. Finally $\hat{\alpha}$ is obtained from $\hat{\alpha} = \bar{Y} - \bar{X}_1\hat{\theta}_1 - \ldots - \bar{X}_p\hat{\theta}_p$.

APPENDIX B: COMPUTATIONAL DETAILS OF REGRX ALGORITHM[2]

The Gaussian elimination algorithm is used to compute the statistics required to implement the stepwise program (Draper & Smith, 1966; Efroymson, 1960). The algorithm depends on the following observation. If $\mathbf{P}$ and $\mathbf{Q}$ are nonsingular matrices, then the two systems $\mathbf{R}_{11}\mathbf{b} = \mathbf{g}$ and $\mathbf{PR}_{11}\mathbf{Qh}=\mathbf{Pg}$ are equivalent in the sense that $\mathbf{h}$ is a solution of the second system if and only if $\mathbf{Qh}$ is a solution of the first. Hence, if the second system can be solved for some $\mathbf{P}$ and $\mathbf{Q}$, then the solution of the first system is easily derived. In particular, if $\mathbf{P}$ and $\mathbf{Q}$ are such that $\mathbf{PR}_{11}\mathbf{Q}$ is triangular or diagonal, then the second system can be solved immediately. In practice, $\mathbf{g}$ appears as a subvector of a row and column in a larger matrix $\mathbf{R}$ which also includes $\mathbf{R}_{11}$ as a submatrix.

Any r x c matrix $\mathbf{A}$ may be written in partitioned form as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \cdot \cdot \cdot \cdot \mathbf{A}_{1b} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \cdot \cdot \cdot \cdot \mathbf{A}_{2b} \\ \\ \mathbf{A}_{s1} & \mathbf{A}_{s2} \cdot \cdot \cdot \cdot \mathbf{A}_{sb} \end{bmatrix}$$

where $\mathbf{A}_{ik}$ is r_i x c_k, $\sum_{i=1}^{s} r_i = r$ and $\sum_{k=1}^{h} c_k = c$

A typical partitioning of $\mathbf{R}$ is the following $\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{g} \\ \mathbf{g}^T & 1 \end{bmatrix}$

In the REGRX algorithm, as in most regression algorithms, the (i,j) entry of $\mathbf{R}$ is the Pearson product moment correlation coefficient between variables i and j based on the sample data for the regression problem. The superscript T denotes that the column vector $\mathbf{g}$ has been transposed into row vector form.

---

[2] See footnote in Appendix A.

In the Gaussian elimination algorithm, $\mathbf{Q}$ is the identity matrix. The matrix $\mathbf{P}$ is obtained as a product of the factors $\mathbf{P}^{(1)}$, $\mathbf{P}^{(2)}$,... At the first step, $\mathbf{P}^{(1)}$ is calculated and $\mathbf{R}$ is transformed to $\mathbf{P}^{(1)}\mathbf{R}$. Renaming $\mathbf{P}^{(1)} = \mathbf{M}^{(1)}$ and $\mathbf{P}^{(1)}\mathbf{R} = \mathbf{R}^{(1)}$, the succeeding steps proceed by calculating $\mathbf{P}^{(i)}$, $\mathbf{M}^{(i)} = \mathbf{P}^{(i)}\mathbf{M}^{(i-1)}$, and $\mathbf{R}^{(i)} = \mathbf{P}^{(i)}\mathbf{R}^{(i-1)}$.

Let i' denote the variable to be added or deleted from the equation for step i-1. If variable i' is added, then $\mathbf{P}^{(i)}$ is computed so that the i' column of $\mathbf{R}^{(i)}$ is the i' column of the identity matrix. If variable i' is deleted, then $\mathbf{P}^{(i)}$ is computed so that the i' column of $\mathbf{M}^{(i)}$ is the i' column of the identity matrix. The matrix $\mathbf{P}^{(i)}$ is equal to the identity matrix except in column i'. The i' column of $\mathbf{P}^{(i)}$ is chosen in the following manner. If variable i' is being added and the i' column of $\mathbf{R}^{(i-1)}$ is denoted by $(a_1, a_2, \ldots, a_{i'}, \ldots, a_v)^T$, then the i' column of $\mathbf{P}^{(i)}$ is $-\frac{1}{a_{i'}}(a_1, \ldots, a_{i'-1}, -1, a_{i'+1}, \ldots a_v)^T$. If variable i' is being deleted and the i' column of $\mathbf{M}^{(i-1)}$ is denoted by $(a_1, a_2, \ldots, a_{i'}, \ldots a_v)^T$, then the i' column of $\mathbf{P}^{(i)}$ is $-\frac{1}{a_{i'}}(a_1, \ldots, a_{i'-1}, -1, a_{i'+1}, \ldots, a_v)^T$. Recalling that L denotes the set of variables in the regression equation for step i and E contains all independent variables that are not in L, it is easy to see that if $j \in L$ (j is an element of L), then column j of $\mathbf{R}^{(i)}$ is equal to column j of the identity matrix; and if $k \in E$, then column k of $\mathbf{M}^{(i)}$ is equal to column k of the identity matrix.

Let p denote the number of elements in L. Symmetrically reorder the rows and columns of $\mathbf{R}^{(i)}$ so that the first p rows and columns of the reordered matrix will coincide with the rows and columns of $\mathbf{R}^{(i)}$ corresponding to the elements of L. Mathematically this is accomplished by postmultiplying $\mathbf{R}^{(i)}$ and $\mathbf{M}^{(i)}$ by a permutation matrix which is denoted by $\mathbf{Q}_L$ and premultiplying $\mathbf{R}^{(i)}\mathbf{Q}_L$ and

$\mathbf{M}^{(i)}\mathbf{Q}_L$ by $\mathbf{Q}_L^T$. Thus the matrices $\mathbf{Q}_L^T\mathbf{R}^{(i)}\mathbf{Q}_L$ and $\mathbf{Q}_L^T\mathbf{M}^{(i)}\mathbf{Q}_L$ will have the special forms

$$(1\text{-}1) \quad \mathbf{Q}_L^T\mathbf{R}^{(i)}\mathbf{Q}_L = \begin{bmatrix} \mathbf{I}_p & \mathbf{D}_{12} \\ \mathbf{0} & \mathbf{D}_{22} \end{bmatrix}$$

where $\mathbf{I}_p$ is the p x p identity matrix

$\mathbf{0}$ is an (m-p) x p matrix of 0s

$\mathbf{D}_{12}$ is p x (m-p)

$\mathbf{D}_{22}$ is (m-p) x (m-p)

$$(1\text{-}2) \quad \mathbf{Q}_L^T\mathbf{M}^{(i)}\mathbf{Q}_L = \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{U} & \mathbf{I}_{m-p} \end{bmatrix}$$

where $\mathbf{S}$ is p x p

$\mathbf{U}$ is $(m\text{-}p)$ x p

$\mathbf{0}$ is a p x (m-p) matrix of 0s

$\mathbf{I}_{m-p}$ is the (m-p) x (m-p) identity matrix

If this same reordering of rows and columns is performed on $\mathbf{R}$, then from

$\mathbf{R}^{(i)} = \mathbf{M}^{(i)}\mathbf{R}$ the following matrix identity must hold.

$$\mathbf{Q}_L^T\mathbf{R}^{(i)}\mathbf{Q}_L = \mathbf{Q}_L^T\mathbf{M}^{(i)}\mathbf{R}\mathbf{Q}_L = \left[\mathbf{Q}_L^T\mathbf{M}^{(i)}\mathbf{Q}_L\right]\left[\mathbf{Q}_L^T\mathbf{R}\mathbf{Q}_L\right] \text{ or}$$

$$(1\text{-}3) \begin{bmatrix} \mathbf{I}_p & \mathbf{D}_{12} \\ \mathbf{0} & \mathbf{D}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{U} & \mathbf{I}_{m-p} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{12}^T & \mathbf{R}_{22} \end{bmatrix}$$

where $\mathbf{Q}_L^T\mathbf{R}\mathbf{Q}_L = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{12}^T & \mathbf{R}_{22} \end{bmatrix}$

Any c x d matrix $\mathbf{B}$ is said to be partitioned conformable to the matrix $\mathbf{A}$ if

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \cdots & \mathbf{B}_{1q} \\ \mathbf{B}_{b1} & \cdots & \mathbf{B}_{bq} \end{bmatrix}$$

where $\mathbf{B}_{kj}$ is $c_k \times d_j$, $\sum_{k=1}^{b} c_k = c$, and $\sum_{j=1}^{q} d_j = d$.

The product $\mathbf{C} = \mathbf{AB}$ may be written in partitioned form as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1q} \\ \mathbf{C}_{s1} & \cdots & \mathbf{C}_{sq} \end{bmatrix}$$

where $\mathbf{C}_{ij} = \sum_{k=1}^{b} \mathbf{A}_{ik} \mathbf{B}_{kj}$ is $r_i \times d_j$

Performing the matrix multiplication and equating corresponding partitions gives

(i) $\mathbf{S} = \mathbf{R}_{11}^{-1}$

(ii) $\mathbf{D}_{12} = \mathbf{R}_{11}^{-1} \mathbf{R}_{12}$

(1-4)

(iii) $\mathbf{U} = -\mathbf{D}_{12}^{T}$

(iv) $\mathbf{D}_{22} = \mathbf{R}_{22} - \mathbf{R}_{12}^{T} \mathbf{R}_{11}^{-1} \mathbf{R}_{12}$

Note that $\mathbf{R}_{11}$ is the correlation matrix of the variables in the set L. Let $\mathbf{g}$ denote the column of $\mathbf{R}_{12}$ corresponding to variable v, where either $v \in E$ or v is the variable number of the response. Let $\mathbf{b}$ denote the corresponding column of $\mathbf{D}_{12}$. The elements of $\mathbf{g}$ are the correlations of variable v with the variables in the set L and 1-4(ii) implies that $\mathbf{b}$ satisfies the equation $\mathbf{R}_{11}\mathbf{b} = \mathbf{g}$. Therefore, the elements of $\mathbf{b}$ are the standardized regression weights for the

regression of variable v on the variables in set L. 1-4(iv) implies that the diagonal entry D_{vv} of $\mathbf{D}_{22}$ directly below the **b** column of $\mathbf{D}_{12}$ is equal to $1 - \mathbf{g}^T \mathbf{R}_{11}^{-1}\mathbf{g} = 1 - \mathbf{g}^T\mathbf{b} = 1 - R^2_L(v)$ where $R^2_L(v)$ is the squared multiple correlation coefficient for the regression of variable v on the set of variables L. The off-diagonal elements of $\mathbf{D}_{22}$ may be converted to partial correlation coefficients after dividing by the square root of the diagonal elements in the same row and column. Thus the (j,k) element of $\mathbf{D}_{22}$ divided by the square roots of the (j,j) and (k,k) elements is the partial correlation coefficient between variables v_1 and v_2 after removing the linear influence of the variables in the set L, where v_1 and v_2 refer to the variables occupying the j and k columns (and rows) of $\mathbf{D}_{22}$.

Further characterizations of the elements of $\mathbf{S}$, $\mathbf{D}_{12}$, and $\mathbf{D}_{22}$ are obtained through a careful study of an individual step in the elimination procedure. Figure B1 is a representation of the operations performed during step i+1 showing the transitions $\mathbf{R}^{(i)}$ to $\mathbf{R}^{(i+1)}$ and $\mathbf{M}^{(i)}$ to $\mathbf{M}^{(i+1)}$ for the case where variable j is deleted from the regression equation. In Figure B1, L denotes the set of p variables in the equation at step i, so $j \in L$. $\mathbf{Q}_L$ denotes a permutation matrix that reorders variables so that all variables in L appear first; moreover, within L, variable j appears last (i.e., p^{th}). Figure B2 shows the same transitions for the case where variable j is added. In Figure B2, L denotes the set of variables in the equation at step i+1, so $j \in L$. $\mathbf{Q}_L$ has the same function as described for Figure B1. It should be mentioned that the variable j referred to in Figure B1 is not the same variable j referred to in Figure B2. Also, the partition components of the matrices appearing in Figure B1 are not the same as the corresponding partition components in Figure B2 although the names used in both figures are the same.

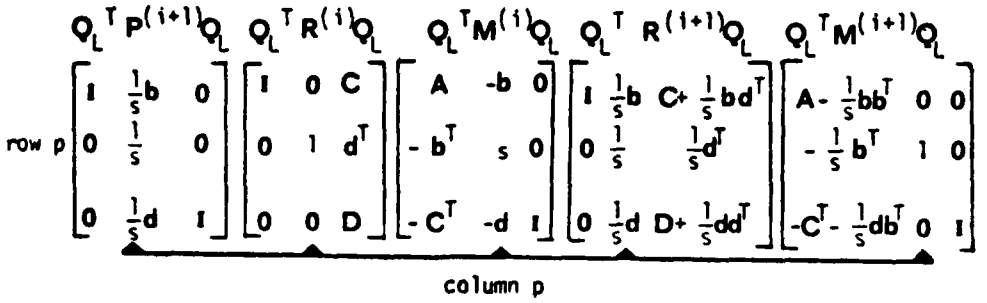


Figure B1. Representation of Matrices Used During Elimination Step i+1 for Deletion of Variable j

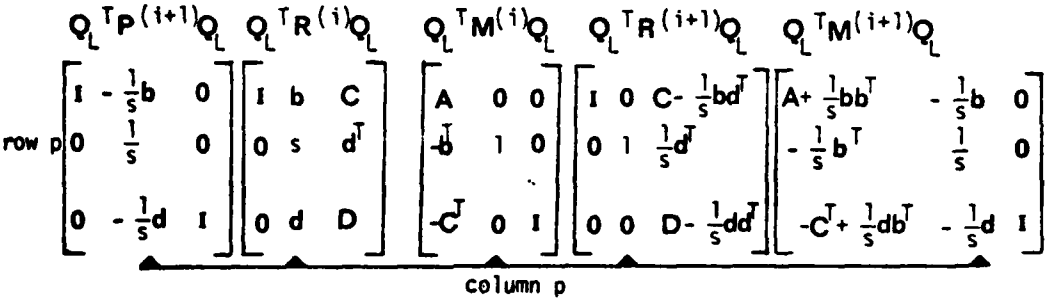


Figure B2. Representation of Matrices Used During Elimination Step i+1 for Addition of Variable j.

Recall that to delete variable j at step i+1 the matrix $\mathbf{P}^{(i+1)}$ is chosen so that column j of $\mathbf{M}^{(i+1)} = \mathbf{P}^{(i+1)}\mathbf{M}^{(i)}$ will be equal to column j of the identity matrix. In Figure B1, the rows and columns of the matrices $\mathbf{P}^{(i+1)}$, $\mathbf{R}^{(i)}$, $\mathbf{R}^{(i+1)}$, $\mathbf{M}^{(i)}$, and $\mathbf{M}^{(i+1)}$ have been reordered by means of the permutation matrix $\mathbf{Q}_L$ for the purpose of simplifying their partitioned form. In the reordered matrices, the elements corresponding to variables in the set L occupy the first p rows and columns and the entries for variable j occupy the p^{th} row and column. Let v denote any variable not in L. Thus, either $v \in E$ or v is the variable number of the response. The component of $\mathbf{d}$ corresponding to variable v is denoted by d_v. Similarly the

diagonal entry of $\mathbf{D}$ corresponding to riable v is denoted by D_{vv}.

Comparing $\mathbf{Q}_L^T\mathbf{R}^{(i)}\mathbf{Q}_L$ with (1-1), and recalling (1-4) (ii) and (iv), it follows that d_v is the standardized regression coefficient $B_{Lj}(v)$ for variable j in the regression of variable v on the set of variables L and that $D_{vv} = 1-R^2_L(v)$. If v is a variable not in the set L, then $B_{Lj}(v)$ denotes the standardized regression coefficient for variable j in the regression of variable v on the set of variables in L, and $\rho_{L-j}(j,v)$ denotes the partial correlation between variables j and v after partialling out tne linear influence of variables in L-j. A similar comparison for

$\mathbf{Q}_L^T\mathbf{R}^{(i+1)}\mathbf{Q}_L$ snows tnat $\frac{1}{s} = 1-R^2_{L-j}(j)$, $\frac{d_v^2}{s} + D_{vv} = 1-R^2_{L-j}(v)$, and $d_v/s = \rho_{L-j}(j,v)\sqrt{\frac{1}{s}\cdot\left(\frac{d_v^2}{s} + D_{vv}\right)}$. From these relationships, tne following results can be derived.

(1) Characterization of the standardized regression coefficient in terms of a partial correlation coefficient.

$$B_{Lj}(v) = \rho_{L-j}(j,v)\sqrt{\frac{1 - R^2_{L-j}(v)}{1 - R^2_{L-j}(j)}} = d_v$$

(2) Characterization of the standardized regressi .cient in terms of tne increase in tne squared multiple ciation coefficient due to tne addition of variable j ' ependent contribution).

$$B^2_{Lj}(v) = \frac{R^2_L(v) - R^2_{L-j}(v)}{1 - R^2_{L-j}(j)}$$

(3) Characterization of the independent contribution of a variable in terms of the partial correlation coefficient.

$$R^2_L(v) - R^2_{L-j}(v) = \rho^2_{L-j}(j,v)\left(1 - R^2_{L-j}(v)\right) = \frac{d_v^2}{s}$$

(4) Characterization of the partial F-statistic for the hypothesis $\beta_{Lj}(v) = 0$.

$$F_j = \frac{R^2_L(v) - R^2_{L-j}(v)}{1 - R^2_L(v)}(n-p-1) = \frac{\rho^2_{L-j}(j,v)}{1 - \rho^2_{L-j}(j,v)}(n-p-1) =$$

$$\frac{B^2_{Lj}(v)}{\left[\dfrac{1 - R^2_L(v)}{1 - R^2_{L-j}(j)}\right]}(n-p-1) = \frac{d_v^2}{sD_{vv}}(n-p-1)$$

Note that this last formula also allows a characterization of the standard error of the standardized regression coefficient. It is known that $F_j = t_j^2 = B^2_{Lj}(v)/s^2_{Lj}(v)$ where t_j is the t-statistic for the hypothesis $\beta_{Lj}(v) = 0$ and $s_{Lj}(v)$ is the standard error of $B_{Lj}(v)$.

(5) Characterization of the standard error of the standardized regression coefficient.

$$s^2_{Lj}(v) = \left[\frac{1 - R^2_L(v)}{1 - R^2_{L-j}(j)}\right]\frac{1}{(n-p-1)} = \frac{sD_{vv}}{n-p-1}$$

In two final results, a characterization for the elements of the inverse and the determinant of the correlation matrix is obtained. Comparing $\mathbf{Q}_L^T \mathbf{M}^{(i)} \mathbf{Q}_L$ with (1-2) (using (1-4)(i)) and denoting the correlation matrix for the variables in the set L by $\mathbf{R}_{11}$, it follows that the diagonal entry, s, of $\mathbf{R}_{11}^{-1}$ corresponding to variable j is equal to the reciprocal of $1- R^2_{L-j}(j)$. Comparing $\mathbf{Q}_L^T \mathbf{R}^{(i+1)} \mathbf{Q}_L$ with (1-1) and recalling (1-4)(ii), it also follows that $\frac{1}{s}\mathbf{b}$ is the vector of standardized regression weights for the regression of variable j on the remaining variables in the set L. Therefore, the non-diagonal entries of the column of $\mathbf{R}_{11}^{-1}$ corresponding to variable j have a simple relation to the standardized regression weights for the regression of variable j on the remaining variables in the set L.

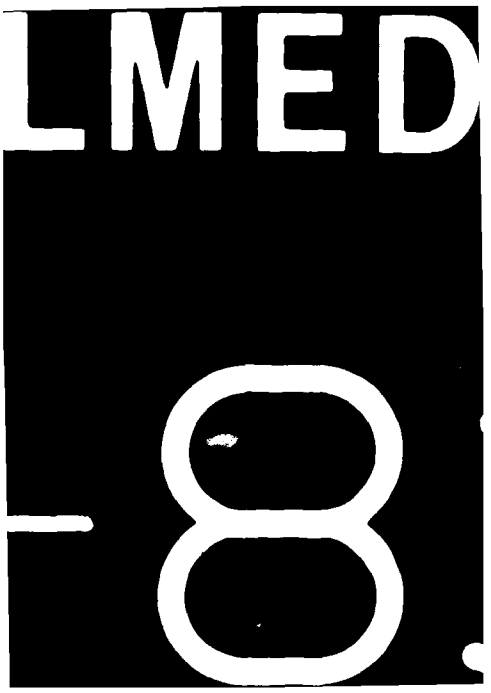
To obtain the expression for the determinant, suppose that $\mathbf{P}^{(i+1)}$, $\mathbf{P}^{(i+2)}$,..., $\mathbf{P}^{(i+p)}$ were chosen to successively delete variables from L until it was empty; then the expression $\mathbf{P}^{(i+p)} \ldots \mathbf{P}^{(i+1)} \mathbf{M}^{(i)} = \mathbf{I}$ would hold. This fact implies that $\det(\mathbf{P}^{(i+p)} \ldots \mathbf{P}^{(i+1)}) = 1/\det(\mathbf{M}^{(i)})$. It is also known that $\det(\mathbf{M}^{(i)}) = \det(\mathbf{R}_{11}^{-1}) = 1/\det(\mathbf{R}_{11})$ and $\det(\mathbf{P}^{(i+1)}) = \frac{1}{s} = 1-R^2_{L-J}(j)$. A generalization of this relationship gives $\det(\mathbf{P}^{(i+2)}) = 1-R^2_{L-j-j_2}(j_2)$, ..., $\det(\mathbf{P}^{(i+p-1)}) = 1-R^2_{L-j-j_2-\ldots-j_{p-1}}(j_p-1)$, $\det(\mathbf{P}^{(i+p)}) = 1$, where $j_2, j_3 \ldots, j_p$ is the order of deletion of variables. Simplifying this notation gives the following general result:

$$\det(\mathbf{R}_{11}) = (1 - R^2_{2.1})(1-R^2_{3.21})(1-R^2_{4.321}) \ldots (1-R^2_{p.p-1\ldots1})$$

where, for example, $R^2_{4.321}$ represents the squared multiple correlation coefficient from the regression of variable 4 on variables 1 through 3.

The formula for the partial F for entry remains to be derived. This is most easily accomplished by means of Figure B2. As before, let d_v denote the component of $\mathbf{d}$ corresponding to variable v and let D_{vv} denote the diagonal entry of $\mathbf{D}$ corresponding to variable v. Comparing $\mathbf{Q}_L^T\mathbf{R}^{(i)}\mathbf{Q}_L$ and $\mathbf{Q}_L^T\mathbf{R}^{(i+1)}\mathbf{Q}_L$ with (1-1) and recalling (1-4)(iv), it follows that $1 - R^2_{L-j}(v) = D_{vv}$, $s = 1 - R^2_{L-j}(j)$, and $1-R^2_L(v) = D_{vv} - \frac{d_v^2}{s}$. Therefore, the increase in the squared multiple correlation due to the addition of variable j is $R^2_L(v) - R^2_{L-j}(v) = \frac{d_v^2}{s}$. This gives the following computational formula for the partial F-statistic for the entry of variable j.

$$F_j = \frac{R^2_L(v) - R^2_{L-j}(v)}{1 - R^2_L(v)} (n-p-1) = \frac{\frac{d_v^2}{s}}{D_{vv} - \frac{(d_v^2)}{s}} (n-p-1)$$


LMED
8